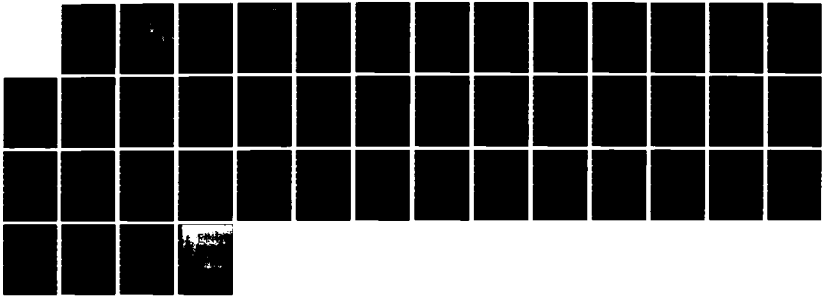
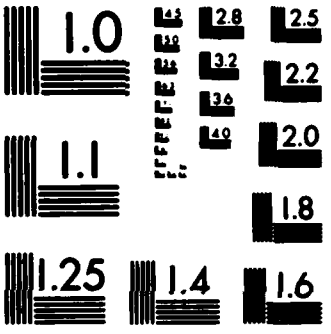
AD-A131 830 RACIAL BIAS AND PREDICTIVE VALIDITY IN TESTING FOR SELECTION(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA R A WEITZMAN JUL 83 NPS54-83-008 1/1

UNCLASSIFIED F/G 5/10 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A131830

NPS 54-83-008

NAVAL POSTGRADUATE SCHOOL

Monterey, California

DTIC
ELECTE
AUG 25 1983
S
A

Racial Bias and Predictive Validity
in Testing for Selection

by

R. A. Weitzman

July, 1983

Approved for public release; distribution unlimited

Prepared for;

Naval Postgraduate School
Monterey, Ca 93940

DTIC FILE COPY

83 08 25 194

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral J. J. Ekelund
Superintendent

D. A. Schrady
Provost

This report received no direct research support.

R. A. Weitzman, Associate Professor

Reviewed by:

Richard S. Elster, Chairman
Dept. of Administrative
Sciences

Released by:

William M. Tolles, Dean of Research

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS 54-83-008	2. GOVT ACCESSION NO. AD-A131 85	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Racial Bias and Predictive Validity in Testing for Selection		5. TYPE OF REPORT & PERIOD COVERED Technical, July 1983 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) R. A. Weitzman		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, ca 93940		12. REPORT DATE July 1983 13. NUMBER OF PAGES 37
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Predictive Validity; Test Bias; Selection; Aptitude Tests; Scholastic Aptitude Test; Racial Bias

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In contrast to the Cleary-McNemar view affirmed by Cole in the October 1981 issue of the American Psychologist on testing--"questions of bias are fundamentally questions of validity"--this report shows that freedom from statutory test bias, as interpreted by the courts, is different from predictive validity. Use of a score-adjustment formula developed here to correct for statutory test bias shows in typical cases not only that the correction tends only negligibly to reduce predictive validity but also that the enhancement of predictive

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED

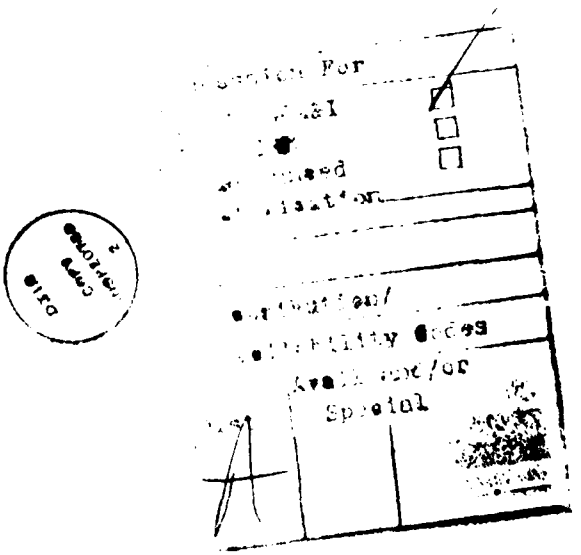
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 Continued

validity without regard to statutory test bias can add a sizable criterion-independent decrement selectively to the already low test scores of low-scoring demographic groups.

Abstract

In contrast to the Cleary-McNemar view affirmed by Cole in the October 1981 issue of the American Psychologist on testing--"questions of bias are fundamentally questions of validity"--this report shows that freedom from statutory test bias, as interpreted by the courts, is different from predictive validity. Use of a score-adjustment formula developed here to correct for statutory test bias shows in typical cases not only that the correction tends only negligibly to reduce predictive validity but also that the enhancement of predictive validity without regard to statutory test bias can add a sizable criterion-independent decrement selectively to the already low test scores of low-scoring demographic groups.



Racial Bias and Predictive Validity in Testing for Selection

Widely perceived as gatekeepers of opportunity, tests have been a popular target of attack in the fight against racial discrimination (e.g., Notes 1 and 2). Mounting challenges have shaken test experts from the complacent position that tests are color-blind measuring instruments no more to blame for abnormally low scores than thermometers are for abnormally high temperatures (Marston, 1971; Weitzman, 1972). Pioneered principally by Guion (1966) and Cleary (1968), recent analyses have revealed occurrences of putative discrimination in which tests often play a leading if unwitting and innocent role. Definitions of fairness in test use have not been uniform, a circumstance to which Flaugher (1978) has drawn particular attention, and treatments intended to assure one form of fairness would seem to work against another. Overcoming discrimination in test use, however, depends on the reconciliation of these differences.

Development of expertise in a technical field like psychological testing tends to produce increasing expectation and tolerance of complication. Whereas the public at large might condemn as biased a test on which white and black people have different means, a test expert is likely to consider this judgment to be premature. In the expert's view, more may be involved than simply a difference in test means. Particularly if the use of the test is to select applicants for work or school, the final verdict on test bias must also take into account subsequent performance on the job or in the classroom. If the racial group

having the higher test mean tends to perform correspondingly better at work or school, then the difference in means may be a more accurate reflection of test validity than of test bias.

Different from most definitions of test bias reported in the literature on the use of tests in selection, the definition adopted in this report contrasts test bias with attenuation of both predictive validity and fairness in test use. Whereas a valid test tends neither to under- nor to overestimate the ability of any racial group, a biased test, as defined here, tends to distinguish between racial groups of equal ability. Because in selection the purpose of a test is prediction rather than measurement, "ability" in this definition refers not to a latent trait, like intelligence, but to the manifest criterion performance to be predicted. Defined in relation to validity, test bias *is*, like validity, a technical property of a test different from fairness, which is a property of test use. Even though a test is free from bias, therefore, a user of the test may still perceive a need for special selection procedures to assure its fair use. Jensen (1980) makes a corresponding distinction between freedom from test bias and fairness in test use, but for him test bias is differential validity, not a tendency to distinguish between racial groups of equal ability. Most other definitions of test bias in the literature turn out on consideration of the distinction between fair and unbiased testing actually to be definitions of fairness in test use.

Procedures for assuring the fair use of specific tests require the use of either multiple cutting scores or equivalent score adjust-

ments. Different definitions of fairness lead to correspondingly different pairs of cutting scores. The next section presents a critique of the use of multiple cutting scores. Proponents of one definition of fairness tend to argue against others. Succeeding sections consider some of the more critical of these arguments, particularly in relation to the definiton of test bias adopted here. The final sections discuss the use of this definiton to approximate bias-free testing both with and without the use of score adjustments.

Test bias can result in the evaluation of the members of one group differently from the members of another group solely because of group membership. The groups may differ with respect to any of a number of demographic variables such as race and sex. To simplify the discussion, this report will refer only to the variable race. Everything said here, however, will apply equally well to other demographic variables that distinguish groups.

The discussion will frequently involve correlations between race and predictor or criterion variables. Whether these correlations are positive or negative depends on the mean scores of the two racial groups on these variables. In accordance with custom, a correlation will be positive if the mean score is higher and negative if the mean score is lower for the traditionally favored group.

Selection, of course, involves both predictor and criterion variables. The problem of concern here is possible predictor bias. Criterion bias, also possible and certainly no less important (Flaugher, 1978; Green, 1975; Gulliksen, Note 3), is beyond the scope of this report.

Multiple Cutting Scores

Multiple-cutting-score definitions of fairness in test use

reflect a variety of standards of fairness. According to Thorndike's (1971) definition, appropriately different cutting scores should yield selection proportions that match success proportions for the different groups. If X per cent of all applicants in Group A would be successful if selected, then the cutting score for Group A should yield the selection of X per cent of all Group A applicants. Cleary's (1968) definition likewise implies the use of different cutting scores if Group A and Group B applicants having the same predictor score have different mean criterion scores. The different cutting scores in this case should correspond to the criterion score separating success from failure.

The Thorndike and Cleary proposals, contrasted at length by Schmidt and Hunter (1974), are only two among many. Linn (1973) and Cole (1973) proposed complementary standards of fairness for different racial groups: equal proportions of successful applicants among the selectees (Linn) and equal proportions of selectees among the successful applicants (Cole). Resembling Linn's, a standard proposed by Einhorn and Bass (1971) is that at their cutting scores members of the different racial groups have equal probabilities of success or failure (risk). Two proposals that directly require only a single cutting score indirectly, through score adjustment, require one for each racial group. These are proposals by Darlington (1971) and McNemar (1975). Darlington, with only a single cutting score, achieves the effect of multiple cutting scores by adjusting the obtained criterion scores; McNemar, similarly, uses race as a predictor

along with the predictor test to form a multiple predictor--the single multiple-predictor cutting score yields the same selection as the multiple cutting scores in the Cleary procedure (see Figure 1). Except for the McNemar and the Cleary procedures, which both use race to maximize validity, all these different procedures involve different standards of fairness and yield correspondingly different cutting scores.

These differences constitute a compelling argument against the possibility of distributional justice (unconditionally fair reward). According to Kaufmann (1973), distributional justice is impossible because of the multiplicity of dimensions of fairness. A single judgment cannot be fair with respect to all dimensions.

Multiple cutting scores, of course, reflect a concern with validity and standards of fairness lacking in the simple use of quotas. As just noted, however, the results of this concern have been unfortunate. Except for the Cleary and McNemar standard, which yields multiple cutting scores that generally favor white over black applicants (Cleary, 1968; Temp, Note 4; Schmidt and Hunter, 1974), all standards of fairness are inconsistent with maximal validity, as all, with no exception, are inconsistent among themselves. The effective difference between quotas and multiple cutting scores may thus be no more than that multiple cutting scores are less predictable than quotas in their impact on selection.

Even proposals to achieve a social consensus regarding the establishment of multiple cuttings scores (e.g., Darlington, 1971; Gross and Su, 1975; and Petersen and Novick, 1976) must fail, though not for technical reasons. Social consensus is subject to test by the courts, and the courts have already cast a

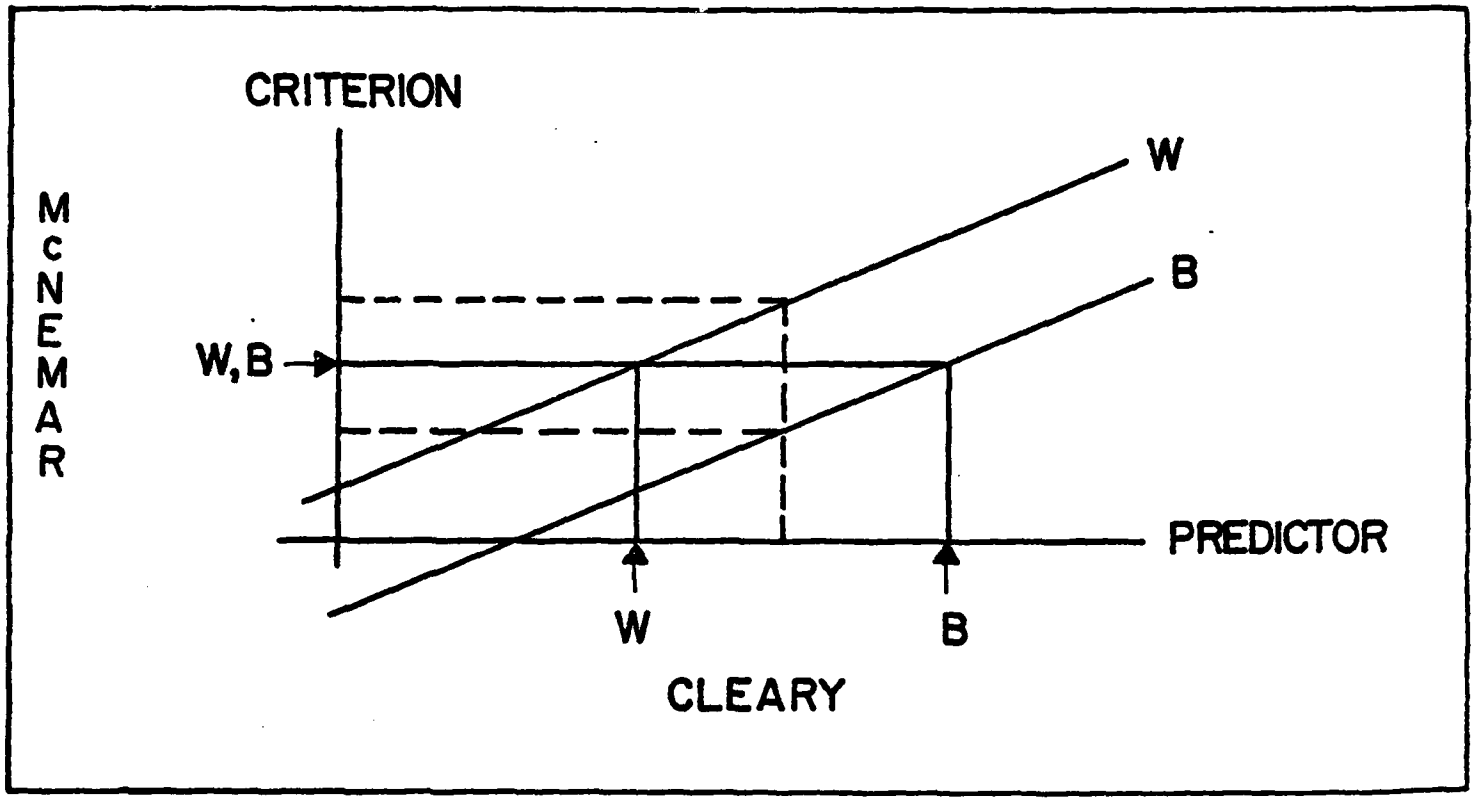


Figure 1. An applicant having the predictor score at the foot of the vertical dashed line will have the same selection fate through the use of either the Cleary or the McNemar procedure--the applicant will be accepted if white (W), rejected if black (B).

shadow over the use of multiple cutting scores. The United States Supreme Court in the pivotal Bakke case[1] rejected the implementation of racial quotas in college-admissions procedures. Succeeding court rulings are likely to move even further in the same direction, and the distance between quotas and multiple cutting scores, as just noted, is not great. Subsequent to the Bakke case, the 3rd District Court of Appeal in California, in fact, ruled against the use of score adjustments to offset lower minority grades or test scores.[2] An admissions officer who adds X points to the score of a minority applicant or uses for the applicant a cutting score X points lower than for a non-minority applicant will, of course, arrive at the same selection result. The California decision thus extends the Bakke rejection of quotas to prohibit the use of multiple cutting scores in selection. The use of any multiplicity of cutting scores, however determined, is indeed interpretable as the implementation of quotas, for is it not the intended effect of this use always to increase the admission proportion of one race relative to another? Even if the intent were separable from the effect, the use of multiple cutting scores would still replace arguable discrimination by unarguable reverse discrimination: the admission of black applicants who score lower than unadmitted white applicants. Test experts can argue that discrimination in one form or the other might occur without the use of multiple cutting scores,

---

[1]Regents of the University of California vs. Bakke, 98 S.Ct. 2733 (1978).

[2]DeRonde vs. Regents of the University of California, 101 Cal. App. 34rd 191 (1980).

but this argument is likely to leave many non-experts unconvinced. Multiple cutting scores would thus appear to constitute a double standard ultimately justifiable only by the claim that the utility of a college education differs for different racial groups. People who are able to agree that a college education has greater utility for one racial group than for another ought also to be able to agree that the ownership of a new automobile has the same relative utilities for the two racial groups. The first agreement entails appropriately different cutting scores for admission to college, however, only if the second entails correspondingly different prices of new automobiles. Social consensus cannot establish multiple cutting scores because--if for no other reason--it points to the use of single ones.

No imposed balance of the future against the past can be just. The past cannot be undone--people who suffered in the past suffered no less even if their descendants fare better. A person does not remove the bias from a coin that has turned up five successive heads by assuring that the next five tosses will produce tails. If each toss is bias-free, however, the proportion of heads will tend to equal the proportion of tails in the long-run. The proposal here is thus not to attempt to balance bias against counter-bias--discrimination against reverse discrimination--but rather to make each test use as nearly bias-free as possible.

The Measurement of Test Bias

A proper definition of test bias ought to imply actions that do not reduce discrimination at the expense of reverse

discrimination. Rather than providing guidance only for the determination of multiple cutting scores or score adjustments, such a definition ought also to constitute the basis for measuring test bias. Armed with a measure of test bias, a test user can identify the test having the least bias among tests that are equally desirable in other respects.

Test bias, as defined here, occurs when, and only when, race (R) accounts for variation on the predictor (P) that has no counterpart on the criterion (C). The occurrence of test bias thus depends on the correlation between race and the component of the predictor uncorrelated with the criterion. This is the part correlation

$$r_{R(P \cdot C)} = \frac{r_{RP} - r_{PC}r_{RC}}{\sqrt{1 - r_{PC}^2}} \qquad (1)$$

Since this correlation is zero when, and only when, the component of P uncorrelated with C is also uncorrelated with R, the inequality

$$r_{R(P \cdot C)} \neq 0 \qquad (2)$$

must define test bias.

This definition of test bias conforms to the requirements of the Civil Rights Act of 1964 as interpreted by the United States Supreme Court in 1971: Either (a) the use of tests in selection must have a numerically <u>equal</u> <u>impact</u> on minority and non-minority applicant groups or (b) any numerical advantage of one group over another must empirically reflect a corresponding <u>job-related</u> advantage.[3] The second of these conditions fails

[3]Griggs vs. Duke Power Co., 91 S.Ct. 849 (1971).

technically whenever a nonzero correlation exists between race and the component of the predictor that is not correlated with the criterion (job), that is, whenever $r_{R(P \cdot C)} \neq 0$.

Not only does a non-zero value of $r_{R(P \cdot C)}$ define test bias, but also any value of $r_{R(P \cdot C)}$ constitutes a measure of test bias. Values of $r_{R(P \cdot C)}$ far from zero represent greater absolute bias than values close to zero. Positive values of $r_{R(P \cdot C)}$ represent bias, negative values counterbias. A test for which $r_{R(P \cdot C)} = 0$ is bias-free.

The definition $r_{R(P \cdot C)} \neq 0$ corresponds closely to a definition proposed by Darlington (1971):

$$r_{RP \cdot C} \neq 0 \ . \tag{3}$$

Stating that at every value of the criterion the mean predictor score tends to be larger for one race than for the other, this definition seems to capture the essence of test bias--a uniform tendency to score one race above the other where no criterion difference exists. The partial correlation in this definition is proportional to the part correlation $r_{R(P \cdot C)}$, with $K = (1 - r_{RC}^{2})^{-\frac{1}{2}}$ as the constant of proportionality:

$$r_{RP \cdot C} = K r_{R(P \cdot C)} \tag{4}$$

Since K does not depend on the predictor, the inequality $r_{RP \cdot C} \neq 0$ constitutes a definition of test bias equivalent to $r_{R(P \cdot C)} \neq 0$. Citing the definition $r_{RP \cdot C} \neq 0$ as the third of four possibilities, Darlington contrasted it with the first, also involving a partial correlation:

$$r_{RC \cdot P} \neq 0 \ , \tag{5}$$

which is, in fact, a formulaic version of the Cleary (1968) definnition, whose use to determine multiple cutting scores was noted earlier. The differences between these two partial-correlation definitions support the adoption of $r_{RP \cdot C} \neq 0$ (or, equivalently, $r_{R(P \cdot C)} \neq 0$) as the proper definition of test bias. The next section examines these differences.

Partial-correlation Definitions

One argument against the use of multiple cutting scores other than Cleary's is that their use attenuates validity (Darlington, 1976). Cleary's exception makes comparison of $r_{RC \cdot P} \neq 0$ with $r_{RP \cdot C} \neq 0$ especially important. To facilitate discussion, $r_{RC \cdot P} \neq 0$ will be called the Cleary definition and $r_{RP \cdot C} \neq 0$ the Darlington 3 definition of test bias. Both involve partial correlations in identical relationships. Their formal resemblance, however, belies fundamental differences. As Hunter and Schmidt (1976) have pointed out, the two definitions are inconsistent. Since $r_{RP \cdot C}$ is proportional to $r_{RP} - r_{PC}r_{RC}$ and $r_{RC \cdot P}$ is proportional to $r_{RC} - r_{PC}r_{RP}$, both cannot be equal to zero at the same time unless $r_{PC} = 1$, which is unlikely to the point of impossibility (no test has perfect validity). This inconsistency has implications for both the definition and the measurement of test bias: If $r_{RP \cdot C} \neq 0$ defines test bias, $r_{RC \cdot P} \neq 0$ cannot; if $r_{RP \cdot C}$ measures test bias, $r_{RC \cdot P}$ must measure something else.

A scatterplot showing the relationship between standardized predictor and criterion measurements ($r_{PC} > 0$) provides a geometric representation of this inconsistency (see Figure 2). If

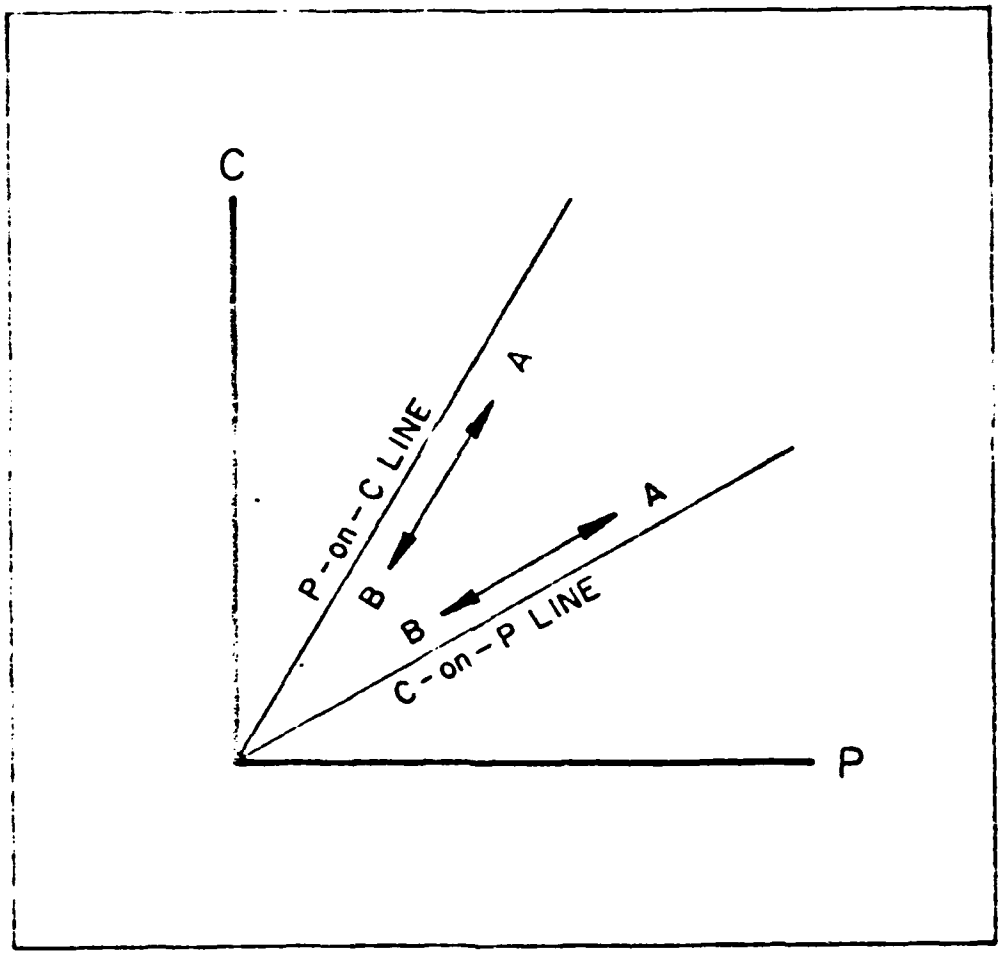


Figure 2. Separation of racial (R) groups, A and B, with respect to two regression lines describing the relationship between standardized predictor (P) and criterion (C) variables when $0 < r_{PC} < 1$. The separation is parallel to the P-on-C line for a bias-free test ($r_{RP \cdot C} = 0$), to the C-on-P line for a test that satisfies the Cleary condition ($r_{RC \cdot P} = 0$) .

$r_{PC} < 1$, the regression lines for P on C and C on P cross at the origin. In the upper-right quadrant (shown), the P-on-C line is above the C-on-P line. In the case of no racial differences, the population means of Groups A and B are equal on each variable. To satisfy the $r_{RP \cdot C} = 0$ (Darlington 3) condition for a bias-free test, racial differences must correspond to the separation of Group A and Group B points parallel to the P-on-C line. The population means for the two groups will now differ on each variable, but not on P for each sub-population having the same value of C. Satisfaction of the $r_{RC \cdot P} = 0$ (Cleary) condition for a bias-free test *corresponds* to the separation of Group A and Group B points parallel to the C-on-P line. Since when $r_{PC} < 1$ this line is not parallel to the P-on-C line, satisfaction of the two conditions cannot occur simultaneously unless $r_{PC} = 1$.

The inconsistency just demonstrated means that any procedure that moves one of the two partial correlations toward zero must simultaneously move the other one away from zero to the extent that r_{PC}^2 differs from one. Procedures that move the Cleary partial correlation toward zero enhance validity (Darlington, 1976). Validity must thus, to some extent, be a casualty of the adoption of Darlington 3 over Cleary.

Technical grounds for deciding in favor of Darlington 3 over Cleary do exist, however. In contrasting the $r_{RC \cdot P} \neq 0$

and the $r_{RP \cdot C} \neq 0$ definitions of test bias, Darlington (1971) showed that the race-predictor correlation (r_{RP}) is considerably greater when $r_{RC \cdot P} = 0$ than when $r_{RP \cdot C} = 0$, particularly for low test validities. This difference in r_{RP} values is due to the correlation between race and the component of the predictor uncorrelated with the criterion ($r_{R(P \cdot C)}$), which satisfaction of the Cleary equality ($r_{RC \cdot P} = 0$) tends to inflate along with $r_{RP \cdot C}$. Attempts to satisfy the Cleary equality thus work to increase the race-predictor correlation inordinately in the process of extracting the entire potential contribution of race to validity. Indeed, if the validity achieved in this process is equal to the race-criterion correlation, whatever its value, then, because r_{RC} must be equal to $r_{PC}r_{RP}$ to satisfy the Cleary equality, the race-predictor correlation will swell to one! The Cleary procedure and its McNemar equivalent have no safeguard to prevent the predictor from becoming race itself ("white, you're in; black, you're out").

The technical decision between the $r_{RC \cdot P}$ and $r_{RP \cdot C}$ definitions of test bias is reducible to a more publically accessible decision. Only when $r_{RC \cdot P}$ is equal to zero does race affect the criterion through the predictor alone. This is certainly an advantage of $r_{RC \cdot P}$. What is wrong, however, if race has an effect on the criterion other than through the predictor? Some

(but not all) validity due to race may be lost, but a zero value of $r_{RC \cdot P}$ permits a perhaps greater danger: Race may have an effect--a possibly large effect--on the predictor other than through the criterion. This effect is bias, which only a zero value of $r_{RP \cdot C}$ can eliminate. The decision between $r_{RC \cdot P}$ and $r_{RP \cdot C}$ thus reduces to the decision between maximizing validity and minimizing bias.

Strong support thus exists for the adoption of the Darlington 3 in preference to the Cleary definition of test bias. Readers familiar with published critiques that appear to challenge the Darlington 3 definition may not yet feel comfortable with this preference, however. The need now, then, is to consider these critiques.

The Hunter-Schmidt Critique

In reference to a predictor correlated with only one factor of a two-factor criterion, Hunter and Schmidt (1976) assumed this factor to be related to race and considered the two cases in which the other factor either was or was not also related to race. The first case becomes a problem for Darlington 3, according to Hunter and Schmidt, if in selection for college the two factors are uncorrelated and the predictor is a pure measure of academic ability. In this case, which fails to satisfy the Cleary equality, Hunter and Schmidt observed that simply a failure also to satisfy the Darlington 3 equality could attach to a perfect predictor the opprobrium of bias. Failure to satisfy the Cleary equality alone, however, could not only attach to this predictor the same opprobrium

but also the opprobrium of attenuated predictive validity. Since the purpose of a predictor is to predict rather than to measure, in fact, a predictor cannot be perfect for its purpose if the criterion predicted is an impure measure of what the predictor is a pure measure. The validity that counts in selection is predictive, not construct, validity. Though pure, therefore, the predictor in this case is not perfect. In the second case, which satisfies the Cleary but not the Darlington 3 equality, Hunter and Schmidt argued that the choice of a predictor correlated with only the first (racial) factor may reflect merely ignorance of the second, not intentional bias. Effect, rather than intention, is the critical concern, however, and solely the attempt to maximize predictive validity without any intention to create bias may produce this case. Apart from these two cases, Hunter and Schmidt wondered how a criterion, as the control variable in Darlington 3, could cause race to covary with a predictor that necessarily preceded it in time. The use of Darlington 3 to indicate test bias makes no reference to causation, however. Any predictor, whatever it measures or fails to measure and whatever the intention that it do so, thus does indeed deserve the opprobrium of bias if, as indicated by Darlington 3, variation on it contains a component due to race that is absent from variation on the criterion.

The Petersen-Novick Critique

In a nice argument against the possibility of distributional justice, Petersen and Novick (1976) rejected the standards of fairness proposed by Thorndike (1971), Linn (1973), and Cole

(1973) for determining multiple cutting scores. (In the case of Thorndike, the argument actually applies to an extended version requiring equal ratios--not necessarily 1:1--of selection to success proportions for the different groups.) Each of these standards has an equally justifiable converse. No more or less justifiable than the "extended Thorndike" requirement that the proportion selected be the same fraction of the proportion successful for every group, for example, is the converse requirement involving the proportion rejected and the proportion

unsuccessful. The cutting scores determined by the application of a standard and its converse are generally different, however. The Thorndike, Linn, and Cole standards of fairness are thus internally inconsistent.

A casual review of the literature suggests that the Petersen-Novick argument might also apply to the Darlington 3 definition of test bias. Since a test for which $r_{RP \cdot C} = 0$ may not require the use of multiple cutting scores to meet Cole's standard of fairness, Cole (1973) associated her definition of "test bias" with Darlington's $r_{RP \cdot C} \neq 0$. Both Hunter and Schmidt (1976) and Petersen and Novick (1976) acknowledged this association without examining it further. Further examination, however, shows that the Darlington 3 definition lies outside the purview of the Petersen-Novick argument. Only in the case of a binary criterion that dichotomizes the population into a subpopulation of successes and a subpopulation of failures may the Darlington 3 condition $r_{RP \cdot C} = 0$ generally obviate the need for multiple cutting scores to meet Cole's standard of fairness, and the Petersen-Novick argument does not extend to this case. Under the conditions of predictor normality and homoscedasticity for the two racial groups within each criterion-defined subpopulation, in fact, the equality of predictor means implied by the Darlington 3 condition $r_{RP \cdot C} = 0$ in turn implies the simultaneous satisfaction of both the Cole standard of fairness and its converse.

The Darlington Critiques

Although Darlington (1971) proposed the equality $r_{RP \cdot C} = 0$ as a possible definition of "culture fairness," he did not himself endorse this proposal. In the same 1971 article and again later (Darlington, 1976), he criticized all definitions involving formulas as too mechanical. Darlington's own preference was for judgmental methods that reflect the relative importance of criterion performance and reverse discrimination. The principle underlying this preference is that some nonzero amount of reverse discrimination is desirable. The arguments made earlier regarding social consensus apply here. No principle favoring reverse discrimination is likely to prevail in a democratic society against the principle that the only desirable discrimination is zero discrimination.

Darlington (1976) extended his criticism of formulaic definitions of "culture fairness" to include possible conflict with validity, mixture of technical (psychometric) and political arguments, insufficient and low-quality selection of minorities from applicant pools having poor minority representation, and the possible arbitrariness of unfavored-group identification. All but the first of these criticisms apply potentially to any procedure designed to remedy test bias or discrimination in test use. As long as both validity and fair test use are desirable, for example, no procedure can be free of either technical or political arguments. The first criticism, moreover, simply reflects reality: Validity and fair test use are, by most standards of fairness, conflicting objectives. Validity will,

indeed, be a casualty of all procedures, including those endorsed by Darlington, that yield selection results different from the results yielded by the Cleary or McNemar procedures. The apparent validity loss in one example cited by Darlington is serious enough to require special comment, however. In this example, based on empirical data, a white applicant whose predicted criterion percentile is 50 would have the same selection fate as a black applicant whose corresponding percentile is only 1. Darlington indicated neither how he used Darlington 3 to obtain this result nor what the percentile difference might be before its use. Percentile differences, in any case, are not validity coefficients. A proper evaluation of validity loss due to the use of any procedure would have not only to compare validity coefficients determined both before and after the use of the procedure but also to include corresponding results of rival procedures in the comparison. One commendable procedure, particularly, is to use neither multiple cutting scores nor differential score adjustment but the most valid available bias-free test. Regardless of the procedure used, however, test users must be prepared to sacrifice incremental validity bought at the expense of test bias or discrimination.

Correction for Test Bias

Test bias as defined by Cleary (1968) is correctable by the use of either the test (predictor) alone with multiple cutting scores (Cleary, 1968) or the test-race multiple predictor with a single cutting score (McNemar, 1975). The first of these options cannot work directly to make $r_{RP \cdot C}$ equal zero because

two predictor scores, while possibly corresponding by regression to the same criterion score (Cleary's procedure), cannot correspond by regression to the same predictor score. A form of score adjustment, like the second (McNemar) option, can work to make $r_{RP \cdot C}$ equal zero, however. The adjustment

$$P^* = Z_P + \beta Z_R \ , \tag{6}$$

where Z_P is the standardized predictor and Z_R the standardized racial measurement and

$$\beta = - \frac{r_{RP} - r_{RC} r_{PC}}{1 - r_{RC}^2} \ , \tag{7}$$

will, in fact, make $r_{RP^* \cdot C}$ equal zero (see Appendix A for complete derivation). Though analogous to McNemar's multiple predictor,

$$P^\dagger = \beta_P Z_P + \beta_R Z_R \ , \tag{8}$$

this adjustment has the effect of minimizing test bias rather than maximizing test validity.

The next section illustrates the effect of the adjustment P* on test validity.

Trade-off between Test Bias and Validity

A trade-off exists between test bias and validity so that the elimination of bias creates a reduction in validity. An example will illustrate this trade-off. Table 1 (left column) describes a test having a validity (r_{PC}) of .41 maximized by race with a race-predictor correlation (r_{RP}) of .40, corresponding to a separation of 1.25 standard-deviation units (S_P) between black ($\bar{P}_B$) and nonblack ($\bar{P}_W$) means on the predictor:

Table 1

Trade-off between Bias and Validity when $r_{RC \cdot P} = 0$

Test data ($r_{RC \cdot P} = 0$)	Adjustment results ($r_{RP*C} = 0$)	Adjustment
$r_{RC} = .16$	$\beta = -.343$	$P^* = Z_P + \beta Z_R$
$r_{PC} = .41$	$r_{P*C} = .39$	$\beta Z_W = -0.12$
$r_{RP} = .40$	$r_{RP*} = .06$	$\beta Z_B = +0.95$

Note. Z_B (-2.775) is the standardized black and Z_W (+0.35) the standardized nonblack racial measurement.

$$r_{RP} = \frac{(\bar{P}_W - \bar{P}_B)\sqrt{p(1-p)}}{S_P} , \qquad (9)$$

where $p = .112$ (proportion of black people in the population). The remaining entries in Table 1 are determinable from these values of r_{PC} and r_{RP}. Since $r_{RC \cdot P} = 0$ and $r_{RP^* \cdot C} = 0$, the numerators in their formulas are also equal to zero; therefore,

$$r_{RC} = r_{RP} r_{PC} \qquad (10)$$

and

$$r_{RP^*} = r_{RC} r_{P^*C} , \qquad (11)$$

where

$$r_{P^*C} = \frac{r_{PC} + \beta r_{RC}}{\sqrt{1 + 2\beta r_{RP} + \beta^2}} \qquad (12)$$

(see Appendix B for complete derivation). The test is biased: $r_{R(P \cdot C)} = .37$. The entries in the leftmost two columns of Table 1 to be compared are the validities r_{PC} and r_{P^*C} and the race-predictor correlations r_{RP} and r_{RP^*}. The differences between r_{RP} and r_{RP^*} and between r_{PC} and r_{P^*C} in Table 1 indicate that elimination of test bias results in a large reduction in the race-predictor correlation but only a small reduction in validity. Whereas r_{RP}^2 is .16 ($r_{RP} = .40$) with a maximal validity of .41 for the unadjusted predictor (left column), adjustment of the predictor to eliminate bias reduces the square of the race-predictor correlation virtually to zero ($r_{RP^*} = .06$) while reducing the maximal validity by only .02 to .39 (middle column). The trade-off thus involves a much greater change in bias, reflected in the relative r_{RP}^2 and

r_{RP^*} values. than validity.

The amount of adjustment from P to P^*, however, was not small: an increase by almost one standard-deviation unit of each black score coupled with a modest decrease of each white score (right column of Table 1). The direction, no less than the amount of adjustment, is notable. Different from the McNemar (Cleary) adjustment, which would generally favor white applicants, this adjustment favors black applicants. The amount and direction of adjustment together reflect substantial bias in the unadjusted predictor against black applicants.

This bias is disturbing because the example cited may not be atypical. Predictors are common on which white means exceed black means by more or less 1.25 standard-deviation units to produce r_{RP} values of around .40 (see Table 2, based on data reported by Temp, Note 4). The median validity of the Scholastic Aptitude Test is .41 (Note 5, p. 16). Corresponding through Equation (9) to the white-black mean difference of .5 standard-deviation units typical of a performance measure (Hunter, Schmidt, and Rauschenberger, 1977, p. 249), the race-criterion correlation of .16 is the correlation with race that a criterion would have if .41 were the validity maximized by race with an r_{RP} value of .40. Use of race to maximize validity can thus, by its effect on the race-predictor correlation, often produce considerable test bias.

Table 2

White-minus-black SAT Mean Differences in Standard-deviation Units

School	Verbal		Quantitative	
	Difference	Standard Deviation	Difference	Standard Dev.
1	1.11	81	1.34	79
2	1.14	81	1.05	93
3	1.88	77	1.75	81
4	1.46	72	1.56	89
5	1.26	109	1.37	108
6	1.34	73	1.10	91
7	1.56	101	1.50	107
8	1.09	94	1.16	85
9	1.43	96	1.27	99
10	1.61	90	1.91	80
11	0.80	89	0.96	85
12	0.83	80	0.77	75
13	0.66	91	1.19	91

Note. Temp (Note 4, p. 10) reported means and standard deviations for the two racial groups separately. Use of the formulas $\bar{P} = p\bar{P}_B + (1-p)\bar{P}_W$ and $S_p^2 = pS_B^2 + (1-p)S_W^2 + p(\bar{P}_B - \bar{P})^2 + (1-p)(\bar{P}_W - \bar{P})^2$, with $p = .112$, provided the total-group information needed to compute the entries in this table.

The use of race to enhance validity need not be overt or even intentional. Nor does this example mean that the SAT is racially biased--the combination of typical values of r_{PC}, r_{RP}, and r_{RC} may be quite rare, and all these values may be quite different in populations of randomly accepted applicants. Because in $r_{R(P \cdot C)}$ conditioning is on the criterion, use of values of r_{PC} and r_{RC} corrected for attenuation due to criterion unreliability may also be more appropriate than use of the observed values. Investigation of the possibility of bias for the SAT and other standardized tests is important, however. What is particularly disturbing is that inadvertently in the process of test construction race may have frequently in the past contributed substantially to test bias while contributing only modestly to validity.

Recapitulation

Pursuing the distinction between test bias and invalidity examined earlier, the previous two sections presented a correction for test bias and used this correction to examine the effects of both bias reduction on validity and validity enhancement on bias. The use of race to enhance validity can produce a race-predictor correlation that is not only much higher than the race-criterion correlation but also more or less as high as the predictor-criterion correlation (see Table 1). These relationships reflect test bias: a nonzero correlation between race and the component of the predic-

tor uncorrelated with the criterion. Elimination of bias reduces the race-predictor correlation substantially while reducing validity only slightly. Corresponding to the positive race-criterion correlation, a positive race-predictor correlation still exists after the elimination of bias. This correlation, however, separates values indicative of bias above it from values indicative of counter-bias below.

Test Choice versus Score Adjustment

If attempts to maximize validity tend to produce test bias, what objective other than maximal validity should a test developer or test user pursue? Who of these two, moreover, should be the more responsible for the pursuit of this objective? These questions arise from the distinction between bias and invalidity, and the first question particularly constitutes a problem that does not exist in the Cleary-McNemar view of test bias. Just as attempts to increase validity tend also to increase reliability, so in the Cleary-McNemar view do these attempts tend to decrease test bias. The problem created by distinguishing between test bias and invalidity is a problem of multiple objectives.

Resolution of this problem is possible by recasting it so that one objective becomes a constraint while the other remains as the sole objective. In the trade-off between the two objectives, a large change in bias corresponds to a small change in

validity. This imbalance suggests that zero bias should be the constraint. A constraint requires fixation at a specific value, moreover, and no such value for validity (other than one, which is all but impossible) corresponds to the value of zero bias. Test developers or test users should thus attempt to maximize validity under the constraint of zero bias.

Since a correction for bias exists, this constrained maximization is easier for the test user than the test developer. The test user need only adjust the predictor by use of equations (6) and (7). Though easier, however, this may not be the generally better solution to the problem. The adjustment will typically favor black applicants by a substantial amount (see Table 1). Argument that the unadjusted scores favor white applicants by the same amount is not likely to preclude charges of reverse discrimination. The better solution politically, though not practically, may thus be for the test developer to attempt to make zero-biased tests with validities as high as possible. The test user would then have the politically acceptable role of choosing among two or more tests the one having maximal validity and minimal bias.

Toward Bias-free Selection

Ruling out the use of multiple cutting scores or differential score adjustment, a test user may thus be left with the option of trying to choose among tests that differ by varying amounts in bias and validity. The choice between two tests is not difficult if both r_{PC} is closer to one and $r_{R(P \cdot C)}$ is

closer to zero for one than for the other. Difficulty arises when for one test r_{PC} and for the other $r_{R(P \cdot C)}$ has the more favorable value. In this case, the choice depends on how the difference between the two r_{PC} values compares with the difference between the two $r_{R(P \cdot C)}$ values. If one difference is substantially larger than the other, then the test user can base his choice solely on the larger difference. Otherwise, considerations additional to bias and validity must determine the choice.

The availability of more than two tests to choose from will generally facilitate the choice. Whether easy or difficult, however, choosing a test with minimal bias may be only a trivial step toward bias-free selection if even the minimal bias is far from zero. The real challenge is thus not test choice but test development.

The basic units of a test are, of course, its items. A test ought to be bias-free, therefore, to the extent that its items are bias-free. Bias-free items are not, of course, items whose difficulties are equal for the two racial groups. A test composed entirely of such items would have an r_{RP} value of zero, lower than r_{RP} values typical of bias-free tests (see Tables 1 and 3) and thus indicative of counterbias. Freedom from bias must take the criterion into account. Using a definition of item bias like Darlington 3, therefore, Scheuneman (1979) developed a method of item analysis to produce bias-free tests. According to Darlington 3, a test is bias-free if in each subpopulation of individuals having the same criterion score the mean predictor scores are equal for the two racial groups;

according to Scheuneman's definition, an item is bias-free if in each subpopulation of individuals having the same test score the item difficulties are equal for the two racial groups. Although the Darlington 3 definition extends naturally from tests to items, the Cleary definition does not. Extended to items, the Cleary definition makes no sense: An item would be bias-free if in each subpopulation of individuals having the same item score (correct or incorrect) the mean test scores were equal for the two racial groups! Actually, the Scheuneman counterpart of Darlington 3 involves complications because each item contributes not only to the total test bias but also to the total test score. The removal or addition of items to reduce test bias may thus alter the Scheuneman bias of the remaining items. Better than partialing out the total test score from the item-race correlation would be partialing out the criterion from this correlation. Criterion availability ought to be no problem for tests used in selection, and for the development of these tests Scheuneman's work provides a useful guide.

Concluding Remarks

Defining test bias consistently with the 1971 United States Supreme Court ruling regarding discrimination in selection, that individual differences having no effect on criterion performance should also have no effect on predictor performance, this report has demonstrated that considerable test bias, so defined, may result from attempts to maximize predictive validity to values attenuated only slightly by score adjustments that eliminate the bias. In a

previous use of score adjustments, Hunter, Schmidt, and Rauschenberger (1977) compared the trade-off effects of satisfying Cleary and other standards of fair test use, including Darlington 3, on black selection ratios and mean criterion performance of all accepted applicants: The ratios tended generally to be much higher and the performance only somewhat lower for Darlington 3 than for Cleary over the 0-1 range of post-adjustment total-group validities. Invalidating their results in numerical detail, though not overall contour, Hunter et al. mistakenly considered these validities to be equal within-group predictor-criterion correlations (pp. 249, 257), unaffected by score adjustments. Taking $(\bar{C}_W - \bar{C}_B)$ generally to be .5 and finding the corresponding $(\bar{P}_W^* - \bar{P}_B^*)$ to satisfy each condition of fair test use, they further considered the difference between this and its observed counterpart $(\bar{P}_W - \bar{P}_B)$ to be the score adjustment for the condition. Since r_{RP^*} and r_{RC} are point-biserial correlations (see Equation (9)), particularly, the Darlington 3 condition $r_{RP^*} = r_{P^*C} r_{RC}$ implies for standardized P* and C that $(\bar{P}_W^* - \bar{P}_B^*) = r_{P^*C}(\bar{C}_W - \bar{C}_B)$. This is exactly the form of the corresponding Hunter et al. equation (p. 257), presented by them without derivation; differing only is the interpretation of the correlation, seen here clearly to be the post-adjustment total-group validity (r_{P^*C}). As different functions of post-adjustment total-group validities, the score adjustments used by Hunter et al. in the comparisons involving Darlington 3 and Cleary are not determinable from available data, and indeed their use to compare the different methods failed to show for each method the

total-group validity changes from unknown though differing pre-adjustment values to predetermined common post-adjustment values. In any event, the comparisons reported by Hunter et al., despite their numerical errors, generally support the position taken here that little, if any, justification exists for predictors of criterion performance, regardless of their validity, to distinguish among groups of individuals who would tend to perform equally well on the criterion if given the chance.

This position unequivocally refutes the assertion by Cole (1981) that "questions of bias are fundamentally questions of validity." Appearing in a special issue of the American Psychologist on testing, Cole's assertion is an authoritative affirmation of the Cleary-McNemar position on test bias. Predictive validity is a compelling concept. Selection error sits opposite predictive validity on a seesaw; as one goes down, the other goes up. The temptation is thus strong to extend the concept of predictive validity, as Cole did, to include other desirable test attributes, as well. Understanding this temptation may perhaps provide some fortification to resist it.

Reference Notes

1. Chapter 1217 (1978), Postsecondary Education--Standardized Tests. Law added as Chapter 3 to Part 65 of the California Education Code.

2. Chapter 672 (1979), Truth in Testing. Law added as Article 7A to the New York State Education Law.

3. Gulliksen, H. When high validity may indicate a faulty criterion (RM 76-10). Princeton, N.J.: Educational Testing Service, 1976.

4. Temp, G. Validity of the SAT for blacks and whites in thirteen integrated institutions (RB-71-2). Princeton, N.J.: Educational Testing Service, 1971.

5. Test use and validity. Princeton, N.J.: Educational Testing Service, 1980.

References

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cole, N. S. Bias in testing. American Psychologist, 1981, 36, 1067-1077.

Darlington, R. B. Another look at "cultural fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Darlington, R. B. A defense of "rational" personnel selection, and two new methods. Journal of Educational Measurement, 1976, 13, 43-52.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75, 261-269.

Flaugher, R. L. The many definitions of test bias. American Psychologist, 1978, 33, 671-679.

Green, D. R. What does it mean to say a test is unbiased? Education and Urban Society, 1975, 8, 33-52.

Gross, A. L., & Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. Journal of Applied Psychology, 1975, 60, 345-351.

Guion, R. M. Employment tests and discriminatory hiring. Industrial Relations, 1966, 5(2), 20-37.

Hunter, J. E., & Schmidt, F. L. Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 1976, 83, 1053-1071.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 1977, 62, 245-260.

Jensen, A. R. Bias in mental testing. New York: The Free Press, 1980.

Kaufmann, W. Without guilt and justice. New York: Wyden, 1973.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

McNemar, Q. On so-called test bias. American Psychologist, 1975, 30, 848-851.

Marston, A. R. It is time to reconsider the Graduate Record Examination. American Psychologist, 1971, 26, 653-655.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Schmidt, F. C., & Hunter, J. E. Racial and ethnic bias in psychological tests. American Psychologist, 1974, 29, 1-8.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Weitzman, R. A. It is time to re-reconsider the GRE--a reply to Marston. American Psychologist, 1972, 27, 236-238.

Appendix A

This appendix derives an adjustment to the predictor score to make $r_{RP \cdot C}$, the partial correlation between race (R) and the predictor (P) controlling for the criterion (C), equal to zero.

Constituting only a single condition, the equality $r_{RP^* \cdot C} = 0$ can determine only a single constant, β, in the multiple predictor $P^* = Z_P + \beta Z_R$, where the Z's denote standardized measurements for the subscript variables. Since the numerator of $r_{RP^* \cdot C}$ is $r_{RP^*} - r_{RC} r_{P^*C}$, the strategy followed is to equate r_{RP^*} to $r_{RC} r_{P^*C}$ and solve for β:

$$(1/N)\Sigma Z_R \, (Z_P + \beta Z_R) = r_{RC} (1/N)\Sigma Z_C \, (Z_P + \beta Z_R) \quad \text{(A1)}$$

or, on simplification and substitution separately on each side,

$$r_{RP} + \beta = r_{RC} r_{PC} + \beta r_{RC}^2 \quad \text{(A2)}$$

with solution for β yielding

$$\beta = - \frac{r_{RP} - r_{RC} r_{PC}}{1 - r_{RC}^2} \, . \quad \text{(A3)}$$

Appendix B

This appendix develops a formula for the validity of P*, r_{P*C}.

The strategy used is direct simplification and substitution (see Appendix A for notation):

$$r_{P*C} = \left(\frac{1}{N}\right)\Sigma \frac{Z_C(Z_P + \beta Z_R)}{S_P^*} \tag{B1}$$

$$= \frac{r_{PC} + \beta r_{RC}}{S_P^*} \tag{B2}$$

where S_P^* is the standard deviation of P*,

$$S_P^* = \sqrt{(1/N)\Sigma(Z_P + \beta Z_R)^2} \tag{B3}$$

$$= \sqrt{1 + 2\beta r_{RP} + \beta^2}\ , \tag{B4}$$

so that

$$r_{P*C} = \frac{r_{PC} + \beta r_{RC}}{\sqrt{1 + 2\beta r_{RP} + \beta^2}}\ . \tag{B5}$$

DISTRIBUTION LIST

	Number of Copies
Defense Documentation Center Cameron Station, Building T 5010 Duke Street Alexandria, VA 22314	2
Dean of Research Code 012 Naval Postgraduate School Monterey, CA 93940	1
Library (Code 0142) Naval Postgraduate School Monterey, CA 93940	2
Professor B. Bloxom	1
Professor R. Elster	1
Professor M. Eitelberg	1
Professor W. McGarvey	1
Professor T. Swenson	1
Professor G. Thomas	1
Professor R. Weitzman Code 54 Naval Postgraduate School Monterey, CA 93940	20

END

FILMED

9-83

DTIC